

Security Analytics of Network Flow Data of IoT and Mobile Devices (Work-in-progress)

Ashish Kundu¹, Chinmay Kundu², and Karan K. Budhraja³

¹ IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA
akundu@us.ibm.com

² KIIT University, Bhubaneswar, India
ckkundu@gmail.com

³ University of Maryland, Baltimore County, MD, USA
karanb1@umbc.edu

Abstract. Given that security threats and privacy breaches are commonplace today, it is an important problem for one to know whether their device(s) are in a "good state of security", or is there a set of high-risk vulnerabilities that need to be addressed. In this paper, we address this simple yet challenging problem. Instead of gaining white-box access to the device, which offers privacy and other system issues, we rely on network logs and events collected offline as well as in realtime. Our approach is to apply analytics and machine learning for network security analysis as well as analysis of the security of the overall device - apps, the OS and the data on the device. We propose techniques based on analytics in order to determine sensitivity of the device, vulnerability rank of apps and of the device, degree of compromise of apps and of the device, as well as how to define the state of security of the device based on these metrics. Such metrics can be used further in machine learning models in order to predict the users of the device of high risk states, and how to avoid such risks.

1 Introduction

Network flow data may be categorized as encrypted and unencrypted traffic. Encrypted traffic allows for transmission specific information such as TCP/IP headers, session information, and details of the cipher suite being used. Unencrypted traffic allows for transmission specific information such as URL tags, the context, version and name of an application, DUID and UID, location, the name and type of the device, and details of the operating system being used. The various network interactions of a device are summarized in Figure 1.

2 Inferences from Network Logs and Events

The problem addressed by this work is presented using the generation of 4 inferences.

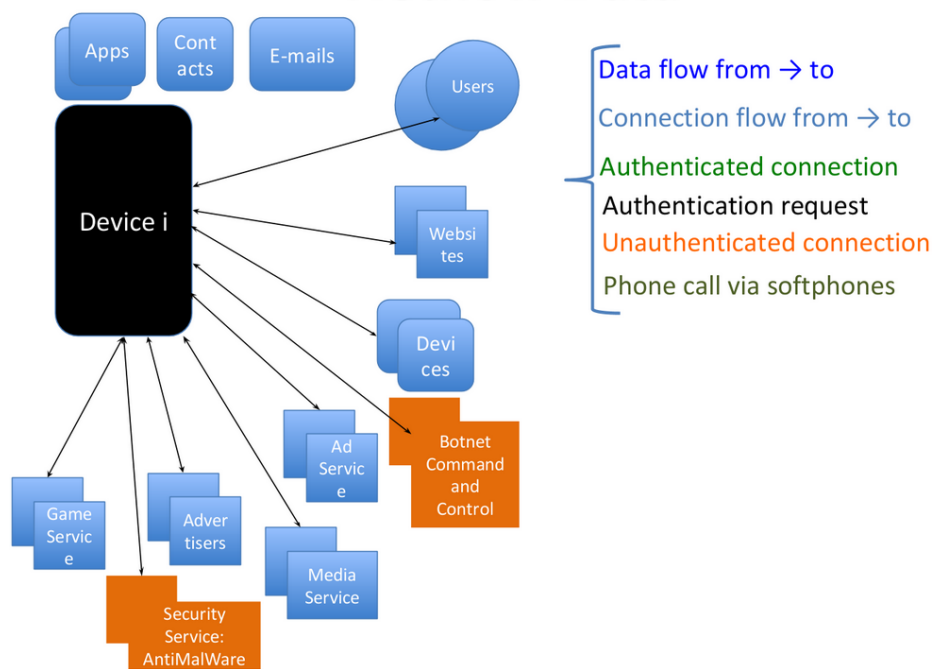


Fig. 1. Summary of network interactions for a network device.

2.1 Inference 1

The first problem is that of inference of details of a network device and the communication associated with it. These details may be summarized as follows.

1. The type of device and its associated software versions.
2. The encryption being used in the traffic associated with the device e.g., DHCP and URL type (e.g., m.google.com, appstore, and other standard IP addresses and URLs being used by current iPhone and Android devices).
3. The information embedded in network packets.
4. The location of a device, observable from its IP address (HTTPS) and HTTP headers.
5. The communication network associated with the device (e.g., services and IP addresses extracted from TCP/IP header information).
6. The frequency of usage of the device and the duration of each session.
7. The advertisements catered to by the device.
8. The categories of applications that are used on the device (e.g., work, games, utility, phone, media, social networks, push type, and pull type).

Application specific information may also be identified. This includes the server addresses used by the application to communicate to a third-party server, if any. This also includes distribution information such as the application version and vendor (may be obtained from application hosting website). Further details include those of device users: the number of users and their identities. Miscellaneous details about the application include the advertisement providers (e.g., Google advertisements) that it communicates with and the means of communication (e.g., HTTP, HTTPS).

This inference also involves the accumulation of network-based statistics. These include the statistics of SSL/TLS used and the various security services used across the network (e.g., communication with Symantec servers). An additional device type may also be obtained for TPM-based services. The statistics also include identification of the type of payload used in encrypted traffic.

2.2 Inference 2

This inference focuses on the communication between devices and applications across the network. The inference involves the generation of user and application profiles corresponding to network data transactions. It also involves identification of the distribution network used by advertisements. These observations can collectively be used to identify a network map of device vulnerability (e.g., the potential behavior of a device with respect to malware propagation across the network, or the observation of botnet components in the network).

2.3 Inference 3

This inference pertains to the compromise of a device and its associated software. An example may be a focus on identification of rootkits in devices. These may

be identified by observation such as the way in which the device behaves on the network, the type of operating system and applications installed, the web addresses that the device communicates to, and a list of potential vulnerabilities.

Hypotheses about an application may be formulated by comparing application behavior with signatures of compromised behavior. This may be performed for identification of applications which are either compromised or exhibit potential of being compromised in the future. This may also identify applications which are currently not compromised but were compromised in the past.

This information allows for the identification of a security lifecycle associated with a network device, where a device moves between protected and compromised states. Arcs in such a state diagram may correspond to details about how the device was cleaned of the compromise e.g., the use of a new installation of the application. It may additionally encode information about the data transmitted during the compromised state.

Analysis of this state diagram allows for identification of applications as malware or re-packaged applications. Such applications exhibit the threat of leaking confidential or private data, user behavior patterns, and sensory data. They may also transmit information related to communication channels used by the device and their associated advertisements and security measures (such as CAPTCHA). The identification of such threats is useful to avoid the compromise of host security and to avoid the propagation of malware.

Finally, such information may be used for forensic and security breach analysis. This involves identification of the types of breaches that may have occurred. This comprises of observation of whether the device is compromised and whether it exhibits anomalous behavior. This translates to servers with which the device communicates and their frequency and duration. Analysis may also involve the observation of an increase in advertisements or the vulnerability of the device. Note that the potential cost of such security breaches is difficult to estimate without the availability of whitebox information for the network. Such availability includes the knowledge of details about the compromise (e.g., the cause and duration of the compromise, and the network actions that were executed by the attacker during that period).

2.4 Inference 4

The final inference to be generated is the energy usage profile of the user. Energy usage allows for a secondary estimate of data transaction by applications (independent applications and the use of web pages). Note that such inference is more useful when the operating system or application being used exhibits adaptive energy consumption (optimizing for lower energy consumption). Since such behavior is integrated in most modern-day software, an energy usage profile is capable of functioning as a secondary source of information on data transactions.

3 Proposed Method

The various utilities of network flow data are summarized in Figure 2. The proposed method then is divided into sections based on the different aspects of analysis of network data.

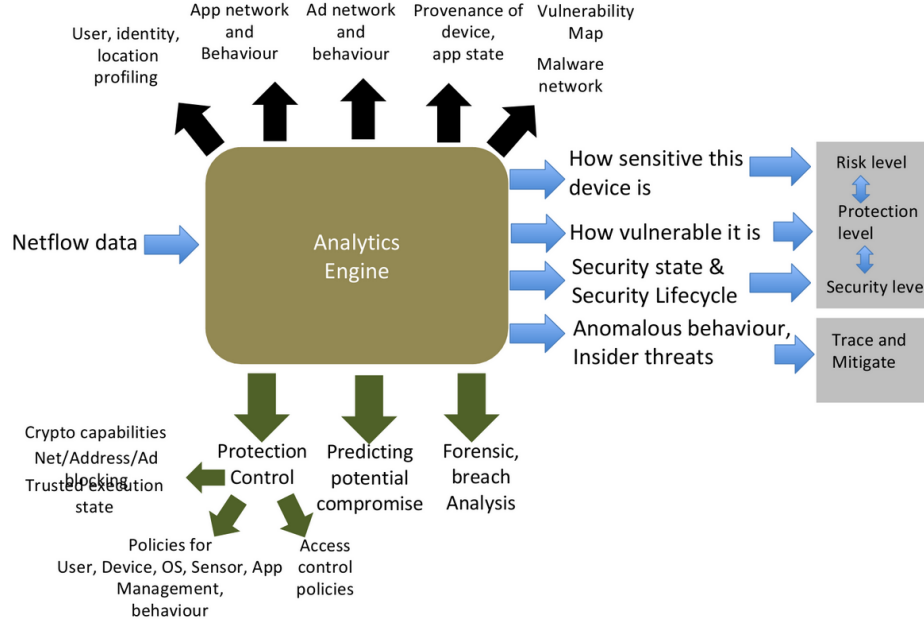


Fig. 2. Using network flow data.

3.1 Sensitivity Analysis

Sensitivity of a device is defined as the importance of data that the device is used with. Specifically, this may be defined as the extent of the device data being personal to the user. The analysis of sensitivity of a device is dependent on the following factors. Note that the analysis is recursive with respect to the sensitivity of the components involved.

1. The sensitivity of the data stored, generated and deleted dynamically (e.g., personal data, passwords and cookies).
2. The sensitivity of the applications used, and data stored by them. Note that the sensitivity of an application may depend on its age.
3. The sensitivity of the other devices and web sites that the user connects to, using this device.

4. The sensitivity of the individuals (other users) that the device is used to contact (e.g., via phone calls, messaging, calls).
5. The sensitivity of the connections and communications associated with the device (e.g., frequency, time, and identity).
6. The amount of data transferred.

3.2 Sensitivity Rank

The sensitivity rank of a device implies the risk level and protection level associated with that device, with respect to sensitivity. Because of the dynamic data and applications associated with the device, sensitivity rank is a dynamic value. For example, a wallet is highly sensitive if it contains more currency, credit cards, personal info, or a combination of these.

Sensitivity rank is computed by building a sensitivity graph. The nodes for this graph correspond to the entities involves i.e., the devices. The edges for this graphs correspond to communication from one entity to another. Sensitivity rank may then be computed as a probabilistic value, similar to PageRank [2]. We therefore define $S(x)$ as the sensitivity rank function for entity x . Similar to PageRank, we have $S(x) \in (0, 1)$.

The following discusses the sensitivity relation between two entities. The sensitivity rank, $S(j)$ of an entity j that entity i communicates with, contributes a weighted value to the sensitivity rank $S(i)$ of device i . A weight function (for an edge connecting entity i and entity j) is defined such that the weight function $W(i, j)$ depends on sensitivity related information such as the frequency of communication, usage, amount of data transferred, how far in the timeline the communication was carried out, age of apps, authenticated or unauthenticated connection. $W(i, j)$ therefore represents weight assigned to the *relative rank* that entity j contributes to $S(i)$.

A high-level recurrence formulation is then presented in Equation 1.

$$S(i) = \sum (W(i, j)S(j)) + S(D(i)) \quad (1)$$

Where a data function, $D(i)$, is included for increased precision. Specifically, $D(i)$ represents the data stored at entity i . This includes the data stored across applications. $D(i)$ is decomposed in Equation 2. $HD(i)$ represents the data generated by and stored at entity i in the past. $CD(i)$ represents the current data generated by and stored at entity i .

$$D(i) = HD(i) + CD(i) \quad (2)$$

Precise computation of sensitivity rank is also dependent on data sources other than those which provide TCP/IP information. This may be latent information learned from network flow data. Note that many applications are identifiable because they do not use TLS. They may also be identified from the advertisements clicked on by the user, or presented as a part of the application.

3.3 Vulnerability Analysis

Vulnerability of a device is defined as the ease with which a device may be compromised. The analysis of vulnerability of a device is dependent on the following factors. Similar to sensitivity, the analysis is recursive with respect to the vulnerability of the components involved.

1. The vulnerability of the applications installed on the device.
2. The vulnerability of the operating system being used by the device.
3. The possibility of vulnerability propagation, i.e. the possibility of a data transaction path existing between a compromised website and vulnerable application on the device.
4. The advertisements associated with applications. This includes the sources of the advertisements and the scripts that they may execute.
5. The information that is transmitted by applications. This comprises of the following.
 - Periodic notifications associated with the application. These include push notifications.
 - The fetching of advertisements and actions associated with clicking on an advertisement.
 - The transmission of sensor-based information. This is significant when considering the possibility of password cracking by the use of sensor readings.
 - The solving of CAPTCHAs [3] associated with the application.

3.4 Vulnerability Rank

Similar to sensitivity rank, the computation of vulnerability rank requires the construction of a graph. A node $(x, V(x))$ in the graph represents entity x with vulnerability $V(x)$, where $V(x)$ is the vulnerability function. An edge from entity x to entity y exists if entity x is compromised by exploiting the vulnerability $V(x)$ and vulnerability $V(y)$. The probability of these vulnerabilities being exploited is represented by the join probability $p((x, V(x)), (y, V(y))) \in (0, 1)$.

The graph is constructed from the known vulnerabilities of different components involved e.g., the operating system, applications, vulnerable websites and services, and advertisements. For an advertisement, the graph may consider how vulnerable or malicious the advertisement network is. The vulnerability of an application depends on the vulnerability of the advertisements that it receives, the source of advertisements, and scripts that are executed thereof. An example of using the graph may be inspection of the existence of a path from a compromised entity to the user's device (paths with $p > 0$). Note that an entity in this graph may be a device, an application, an advertisement, an operating system, or even hardware and firmware components.

As in sensitivity rank, vulnerability rank may then be computed as a probabilistic value, similar to PageRank [2]. We therefore define $V(x)$ as the vulnerability rank function for entity x . Similar to PageRank, we have $V(x) \in (0, 1)$.

The following discusses the vulnerability relation between two entities. The vulnerability rank $V(j)$ of an entity j that entity i interacts with, contributes

a weighted value to the vulnerability rank $V(i)$ of entity i . The weight function $W(i, j)$ depends on $p(x, y)$, the frequency of communication, usage of the entity, the amount of data used, the relative time of occurrence of the communication, and the age of applications associated with the entities. Note that the weight value $W(i, j) \in (0, 1)$.

A high-level recurrence formulation (for the interaction between entity i and entity j) is then presented in Equation 3. This formulation includes Denial of Service (DoS) [1] as a potential threat. Even though an entity may not have any vulnerability (which is highly improbable), $LV(i)$ does not influence the first term in Equation 3, that is due to the remote entities that entity i interacts with.

$$V(i) = \sum (W(i, j)V(j)) + IV(i) + LV(i) \quad (3)$$

Where $IV(i)$ and $LV(i)$ are additional vulnerability functions used for increased precision. $IV(i)$ represents vulnerability due to insiders i.e., the probability that an insider (authorized user with respect to entity i) would become a proponent of compromise of entity i . $LV(i)$ represents the local vulnerability of entity i i.e., the probability that components within entity i can be exploited to produce a compromise.

3.5 Degree of Compromise

The degree of compromise is a composite metric formulated using sensitivity and vulnerability. Degree of compromise, $DC(i)$ of a component i , is based on the following observations in network data.

1. The components and applications specific to a device, that have been compromised.
2. The probability $p(i)$ that the device is compromised based on vulnerability rank $V(i)$ and the network behaviour of the entity i .
3. The criticality of the compromise of entity i , based on the corresponding sensitivity rank $S(i)$.

A high-level recurrence formulation is then presented in Equation 4. j is the component (such as the operating system or application) on the device that has a vulnerability rank $V(j) > 0$.

$$DC(i) = p(i) * S(i) * Sum(DC(j)) \quad (4)$$

3.6 Security State Analysis

The security state of a device can be determined from network data. The formulation of such state dynamics is useful for applications such as risk analysis, taking protective actions in case of a compromise of security, and forensic and security breach analysis. The security state diagram of a device is summarized

in Figure 3. The transitions between different security states over time are summarized in Figure 4. The security state of a network device may be determined by the following.

1. The observation of anomalous behavior with respect to the device.
2. The access of security services (e.g., Symantec servers) by the device.
3. The strength of the cipher suite associated with the device.
4. The protocol used by the device (e.g., HTTP, HTTPS, SRTP, and RTP).

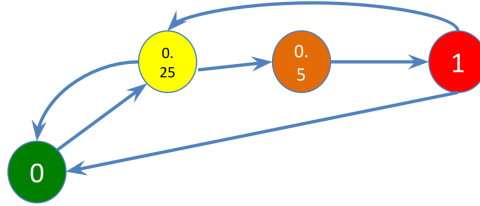


Fig. 3. Security state diagram for a specific device. The diagram and values associated with nodes are dynamic with respect to the applications installed in the device and the network data transactions thereof. The values on nodes represent the degree of compromise.

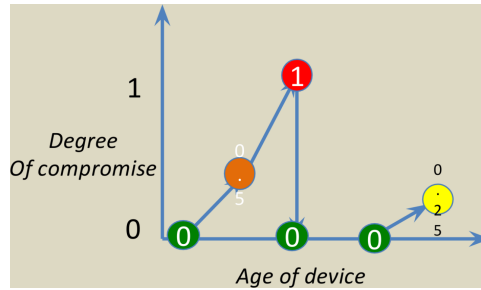


Fig. 4. Security state transitions.

3.7 Protection Based on Network Flow Data Analysis

At a given time, the protection level required to prevent the risk can be associated with the quantification of the risk (involving the computation of $S(i)$, $V(i)$ and $DC(i)$). The details of these computations are presented in Section 3.2, Section 3.4 and Section 3.5. This quantity can then be limited to a threshold. Appropriate actions can then be incorporated across the network. Examples of the different

types of actions that may be executed across the network are enumerated as follows.

1. Strengthen or weaken access control and authorization policies with respect to different network components.
2. Restart a network component. The device may then enter a trusted execution mode, e.g., for financial transactions. This may be required for a network component i that observes $DC(i) > 0$. Further, all future data transactions may be monitored for the given device.
3. Notification of the current security state of a user, and appropriate actions that may be taken.
4. The enforcement of user behavior policies. For example, a user may not be allowed to open a sensitive website such as www.chase.com after it has clicked on an advertisement from low-sensitivity application such as a game (e.g., AngryBirds).
5. The enforcement of mixed user behavior policies with system-driven control.
6. The enabling or disabling of applications, features and sensors across the network based on updated policies.
7. Initiation of backup of device data and the removal of all sensitive data and applications from the device.
8. The blocking of third-party application synchronization (with other devices) for vulnerable applications.
9. The lockdown of network communication at various levels of granularities across the network.
10. Alternative defense mechanisms for the network, in the case where the policy engine governing the network has been compromised.
11. Disabling a device by draining or removal of all associated energy sources. This may be enabled by a remote protection unit that sends targeted scripts via advertisements and web pages to drain energy. For example, if the device is being used as a bot, and the network infrastructure is beyond control by local protection.

Risk analysis may also be translated to long-term actions such as the following.

1. The identification of requirements for software patches required across different devices and types of patches, and their scheduling. This may be based on the security state analysis of a device.
2. The engineering of applications and software. This may depend on the programming models to be enforced, analysis that is required to be incorporated with the program (such as sensitivity and vulnerability).
3. The devising of a methodology by which applications and software are required to support APIs for usage with trusted local and remote services. This enforces dynamic protection and the ability of take appropriate defense actions.

3.8 Forensic Analysis

Forensic analysis of network data involves the use of provenance data collected by the network. This data comprises of the following.

1. Information about the states of different network components
2. Information about the various ranks (such as sensitivity and vulnerability) and degrees of different network components.
3. Other latent information that may be inferred from data transmission across the network.

Such analysis is useful for automated auditing of the behavior of a device, users and applications that are present in the network. An alert may be issued by a background process that continuously checks whether the degree of compromise crosses a threshold value. This process may then also identify the following information.

1. An analysis of any breaches that may have occurred.
2. The methodology that was adopted to compromise a given device.
3. The lack of protections in context of network security that may be incorporated to avoid such compromise in the future.

Alternatively, forensic analysis may be used to analyze the potential cost of a security breach. This may be based on the following.

1. The end-points that the device is communicating with at the time of compromise.
2. Details about the communication, such as its duration and the frequency of such communication.
3. Categorization of the communication to reflect severity of the potential breach. Examples of categories of communication include monetary, political, social, and private.

4 Machine Learning

We develop a Spark-based machine learning stack for implementation of classification of risk levels of apps and of a device. We plan to develop an SVM-based system, following which we plan to develop a neural-network based model and compare their accuracy versus efficiency in predicting risks as well as classification of apps in the risk lattice. We plan to apply the risk classification and risk prediction for multiple devices together that are in a geolocation or in a network.

5 Conclusions

In this paper, we discussed the problem of determining the state of security of a device – mobile or IoT, using big data analytics and machine learning on network logs and events of such devices. We further outlined a set of steps towards remediation of such issues.

References

1. Needham, R.M.: Denial of service. In: Proceedings of the 1st ACM Conference on Computer and Communications Security. pp. 151–153. ACM (1993)
2. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
3. Von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: Captcha: Using hard ai problems for security. In: International Conference on the Theory and Applications of Cryptographic Techniques. pp. 294–311. Springer (2003)